Spatio-temporal approach for Road Scene Classification for Self-driving Cars

Haider Ali
Virginia Tech

Soban Ali
LUMS

Murtaza Taj
LUMS

ABSTRACT

Artificial Intelligence is changing the way we are doing our daily activities. While the developed countries are moving towards self-driving cars, we still lag behind because of less developed infrastructure and lack of work in our local context. While driving self-driving cars, It is always necessary to define a road network of its surrounding. Moreover, solutions to this problem are important for different driving assistance technologies and can predict individuals behaviour. So, this work will provide a novel approach for recognizing various road scenes to define the road network around a self driving car, through the sequence of images. We developed a Spatio-temporal model that recognizes the scene for self-driving cars using spatial as well as temporal features through the history of image frames. Spatial features are the main features to predict the scene shown in the image frame while temporal features try to ensure consistency with previous image frames, hence giving a better result. We divided our work into following parts. (i) data gathering and cleaning (ii) literature review and analyzing previous approaches (iii) training and testing of CNN based models on spatial features (iv) defining a Spatio-temporal model for better result on sequence of images (v) testing on video dataset. We gathered the data from our local streets by installing a camera on the inside of a car's windscreen, while some of the data is from freely available Canadian streets dataset. Our proposed Spatio-temporal model is able to classify scenes with approximately 90% accuracy and correctly identifying phases of transition between them.

KEYWORDS

Spatio-temporal approach, transfer learning, ImageNet, CNN, ResNet, VGG, Self-Driving Cars

ACM Reference Format:
Haider Ali, Soban Ali, and Murtaza Taj. 2021. Spatio-temporal approach for Road Scene Classification for Self-driving Cars. In , .

1 INTRODUCTION

Modern autonomous systems must take decisions in the context of road scenes. Previous works have always neglected the infrastructure of developing countries. Therefore, for a more generic solution, we used the diverse dataset of both developing and developed countries. We collected our own dataset of 14 types of road scenes in Pakistan (a developing country) and used open source Canadian (developed country) dataset. Road scenes included days and nights of market, bridge, underpass, road intersection, expressway, no intersection and narrow street. To make it more diverse, we also used images with different brightness, contrast, zoom and rotation levels. We first approached this problem with spatial classification using a noise free dataset and transfer learning approaches. Different weights for different classes were used to solve the challenge of imbalanced data. We used VGG16, Inception-ResNet-v2 and ResNet-50 models with pre-trained weights on ImageNet dataset and got accuracies upto 98%. To make it perfect for real scenarios, we had to deal with a dataset having noise and scene transitions. Solving with the same approach was giving random predictions whenever there is a noise and scene transitions in the test videos. Therefore, we used spatio-temporal techniques to solve the issue of scene transitions. Spatial features were used for scene classification while temporal features made use of a time window to ensure the consistency with previous image frames, resulting in accuracies upto 91%.

We divided our work into following parts.

- Data gathering and cleaning
- literature review and analyzing previous approaches
- training and testing of CNN based models on spatial features
- defining a spatio-temporal model for better result on sequence of images
- testing on video dataset.

2 RELATED WORK

After generating the required dataset, we looked into various ways through which we can make a model for classification of various road scenes like flyover, underpass etc. The approaches that are used before are mostly based on spatial features. One research [1] has been done to classify scenes using machine learning based approaches including SVMs, linear SPM and ResFeats. In SPM approach, images are represented as histograms of various partitions of different scales. A matrix is computed using a sparse coding to a descriptor set and then a max pooling function is applied to the resulting matrix. One of the better approaches was ResFeats where the features are extracted from the last convolutional layer of deep residual networks pretrained on ImageNet. This approach gave an overall accuracy of 94.7%. Another research [2] suggests a novel approach using a pre-trained network, AlexNet. It first augments the data according to the parameters of AlexNet, then uses CNN layers to train the network. The overall accuracy for indoor scenes was 92.5% and for outdoor scenes 97.8%. However, again these approaches only consider the spatial features and do not include any temporality, and are not very suitable for sequence of images.

3 DATA GATHERING AND CLEANING

To train a model related to our own country's infrastructure there is a need to collect the data of our local cities, that should be as diverse as possible and should cover most of our country's roads. Right now we only targeted rural areas so the data is mostly based on the streets of Lahore that covers all of the scenarios mentioned

, ,

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00

below. To further improve the training model, the dataset is based on day time and night time images in order to provide a thorough training set to deal with scene classification problems regardless of the time of the day. The classes contained in our dataset are as follows:

- Market
- Underpass
- Road Intersection
- Expressway
- No Intersection (straight road)
- Flyover
- Narrow Street

For all these classes we further had two variants, i.e. two categories, one for daytime and other for night time. This makes the overall trainable classes to be equal to 14. We installed the camera inside the windshield of the car and recorded videos while driving on the roads of Lahore. Then, we extracted relevant images (frames) from the videos in order to generate a dataset. After generating the dataset, we divided that into three exclusive groups as training, validation and testing sets. We ensured exclusivity among groups in order to get the better idea of performance of our model. Then we tested it on live videos as well, that will be discussed later.

4 TRAINING MODELS ON SPATIAL FEATURES

Convolutional Neural Network has performed very well for different computer vision classification problems and has outperformed other previous approaches. Therefore, we specifically looked into it and tried its different variants for building a model for our problem. Since there are already very powerful trained networks available like VGG16, Inception-ResNet and ResNet, thus we used transfer learning approach to train our model and fine tune them in order to get better results. We used VGG16, Inception-ResNet-v2 and ResNet-50 models with pre-trained weights on ImageNet dataset. We used these models to extract features from the image frame and then used a simple dense layer followed by a classification layer. We trained these models using our generated dataset and we also tested them on the testing test set. The accuracy of these models have been shared below. We had around 18088 images overall and we used around 14660 images for training, 1620 for validation and 1808 images for testing set. The number of images for different classes were different, therefore we had to assign different weights to classes while training, according to their number of samples for each class. The distribution of training dataset can be seen in Figure 1 The models were left for training with many epochs until there were no further improvements in the validation loss.

5 TESTING

The set that we used for testing had 1808 images. The images in this set were also distributed in the same proportion as in the training set. The results that we obtained on models based on different architectures are in Figure 2

Class Name	Number of Frames/ Image Samples
Night Market	475
Day Underpass	1683
Day Intersection	1033
Night Expressway	765
Day NoIntersection	1009
Day Expressway	563
Night Underpass	2306
Night NoIntersection	393
Night Intersection	1084
Night Flyover	1457
Day Market	821
Day Flyover	963
Night NarrowStreet	1382
Day NarrowStreet	726

Figure 1: distribution of training dataset

Used Model Architecture	Accuracy
VGG16	95%
Inception Resnet V2	96%
Resnet 50	98%

Figure 2: Results on different Architectures

6 SPATIO-TEMPORAL CLASSIFICATION

The dataset that we used for training and testing of our model had distinct features and differences according to their classes i.e. the dataset was cleaned or without any noise, that is also the reason we obtained these pretty good results through our model. But this is not usually the case when we use a model in the real time scenario. We may come across with abrupt traffic change, glare on the camera lens,random passers by specially on narrow streets or market etc and because of these changes we might get continuously changing predictions for different image frames. And that is exactly what happened when we tested our model on videos or in a live road setting. Similar behavior was also seen when a vehicle was changing the scene from one class to the other. While doing testing on videos, we do not want to use all frames from the video. Thus, to increase the visibility of results we extracted 2 frames per second from the video and ran them through our model.

In this particular case of 3, the fluctuations or frequent ups and downs are showing when the vehicle is moving from class 1 to class 6, and then again shifting to class 3. The model got confused even before the actual transition between scenes took place and our model started giving random predictions. In the real world scenario, we do not want this to happen. Thus, we would like to make it a bit more smooth or we do not want our model to get confused when it sees some random shifting among scenes. In order to deal with this problem, we introduce temporality within our classification system. The idea is that whenever a model makes a prediction using one

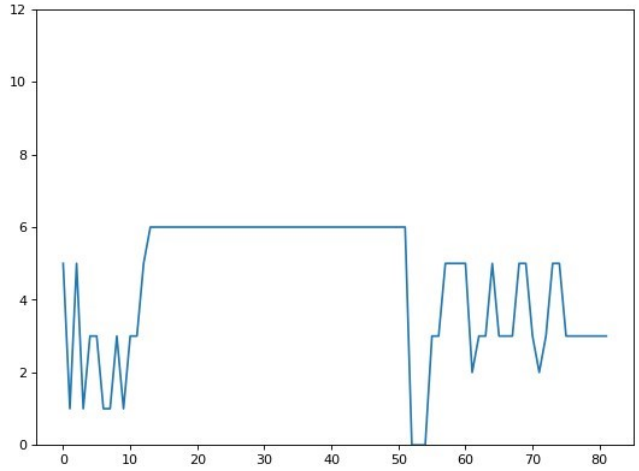

Figure 3: Graph between frames and category - random predictions in test videos - spatial classification approach

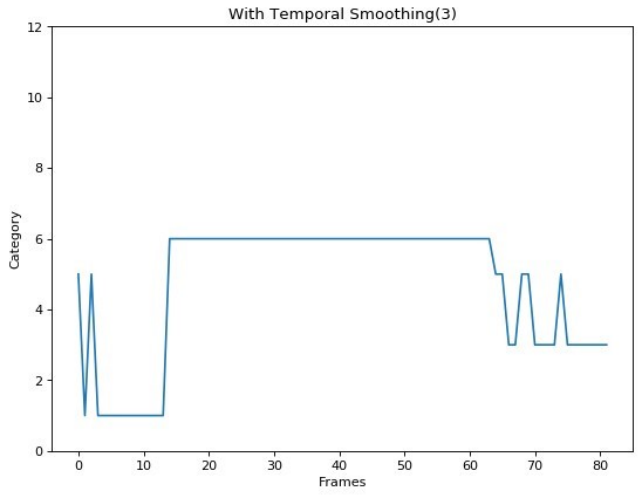

Figure 4: Graph between frames and category of video used in Figure 3 - Spatio-temporal classification approach

frame, we add the probability of some previous image frames to the probability of current image frames. We can change the window size i.e. number of probabilities from previous frames according to our own choice. Greater the size we choose for our image window, more smoothing we will see in the prediction curve. In this way, even if the image frames (or features within them) change abruptly throughout the sequence, the predictions of our model will not be so random or indeterminate. After introducing temporality within our classification system, I ran the same video through the classification model and this time I see the curve in 4.

Similarly, Figure 5 is the scene classification curve from another video, which was mainly the car driving through the market road

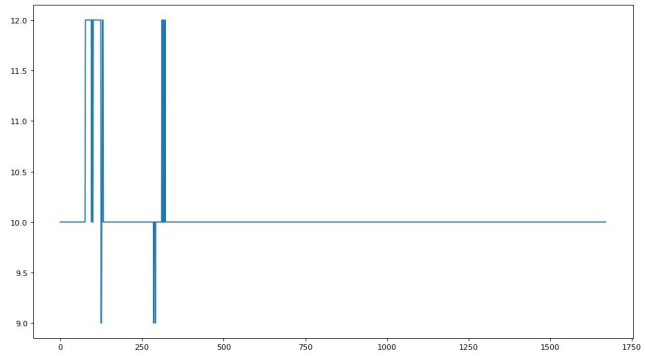

Figure 5: Graph between frames and category of market video - Spatio-temporal classification approach

during night time. Still, we can see random fluctuation or wrong predictions but that can be handled as well if we increase the window size of previous images.

7 DISCUSSION

This project provides an initial work for road scene classification system for autonomous objects, especially in the context of developing countries where the road infrastructure ranges from very good quality to mid-range quality roads like in Old Lahore Markets. Moreover, classification of road scenes can become random for some part of an image sequence because of various reasons, thus our approach will exploit from the temporality of the image sequence to provide a better prediction. However, Performance can be improved further by using more dataset from various other locations and by using multiple camera images on a vehicle. Other spatio-temporal approaches can also be tried in the future

REFERENCES

[1] Himanshu Patel and Hiren Mewada. 2018. Analysis of Machine Learning Based Scene Classification Algorithms and Quantitative Evaluation. *International Journal of Applied Engineering Research* 13, 10 (2018), 7811–7819.

[2] A Yashwanth, Shaik Shammer, R Sairam, and G Chamundeeswari. 2019. A novel approach for indoor-outdoor scene classification using transfer learning. *International Journal of Advance Research, Ideas and Innovations in Technology* 5, 2 (2019), 1756–1762.